ТЕХНОЛОГИЧЕСКИЕ АСПЕКТЫ СБОРА ДАННЫХ СОЦИАЛЬНЫХ СЕТЕЙ*

Митыпов Д. А., Хандаров Ф. В.

dugar.mitypov@gmail.com, fedor.khandarov@gmail.com

Институт экономики и управления БГУ

Исследования в цифровых гуманитарных науках

Улан-Удэ, 27.09.2020

*При поддержке РФФИ, проект 18-312-00186 мол_а

I. ЗАДАЧА СБОРА ДАННЫХ

1. Построение и анализ социального графа

Графовые характеристики Кластеризация Ранжирование (центральности)	Специализированные графовые алгоритмы Оптимизация на графах

2. Построение и анализ таблицы признаков (без графовой информации) => все возможные задачи анализа данных и подходы к их решению

Восстановление регрессии Кластеризация Классификация	Нейронные сети Деревья решений Прочие модели и подходы

3. Вспомогательные задачи распознавания

Распознавание изображений Распознавание видео	Глубокое обучение

II. ПРОБЛЕМЫ СБОРА ДАННЫХ И ИХ РЕШЕНИЯ

1. Возможности, предоставляемые API

API Facebook не позволяет получать данные социального графа

2. Скорость сбора данных

Медленная скорость

3. Достоверность информации и закрытость аккаунтов

Вероятность указания неверной информации

1. Возможности, предоставляемые API

API Facebook не позволяет получать данные социального графа

Selenium WebDriver — инструмент для автоматизации действий веб-браузера

Позволяет автоматизировать действия пользователя социальной сети по открытию страницы и сбору списка друзей, т.о. можно строить социальные графы, в случае, если у социальной сети нет соответствующего функционала в API

Существуют реализации для Python, Scala

2. Скорость сбора данных

Медленная скорость

— Использование GPU для распараллеливания рутинных операций CUDA, OpenCL и др.
#многопоточный сбор социального графа
ОГРАНИЧЕНИЕ: ограничения API по обращению из-под одного аккаунта, или одного IP-адреса — работа с разных IP-адресов и разных аккаунтов
=> организация работы по сети

— Использование возможностей API для пакетных запросов
#за один запрос запросить данных о друзьях не одного, а 25 аккаунтов

3. Достоверность информации и закрытость аккаунтов

Теоретико-вероятностные задачи на:

— определение вероятности факта дружбы для закрытых аккаунтов (по информации об односторонней дружбе у открытых аккаунтов)

— определение вероятности пола, возраста, локации и т.п.

Восстановление данных для закрытых аккаунтов

Проверка данных для фейковых аккаунтов

III. ПРИКЛАДНЫЕ ЗАДАЧИ

1. Социальный граф писателей в сети Facebook

Множество вершин	**Характеристики графа**	**Инструментарий**
Список аккаунтов российских писателей — номинантов и лауреатов литературных премий	Социальный граф Facebook ~ 10^3 вершин ~ 10^4 ребер	Программный комплекс на Python на базе Selenium Webdriver

2. Графы пользователей в сети VK

Множество вершин	**Характеристики графа**	**Инструментарий**
Список аккаунтов российских рок-поэтов Список музыкальных/поэтических сообществ VK	Социальный граф VK ~ 10^3 вершин ~ 10^4 ребер	Программный комплекс на Python на базе VK_API

Спасибо за внимание!